# SPEECH ACTIVITY DETECTION AND SPEAKER DIARIZATION FOR LECTURES

***presented by Xuan Zhu***

RT-06s workshop
Bethesda, Maryland
May 4, 2006

# INTRODUCTION

## Task

- speech activity detection: speech/non-speech
- speaker diarization: who spoke when

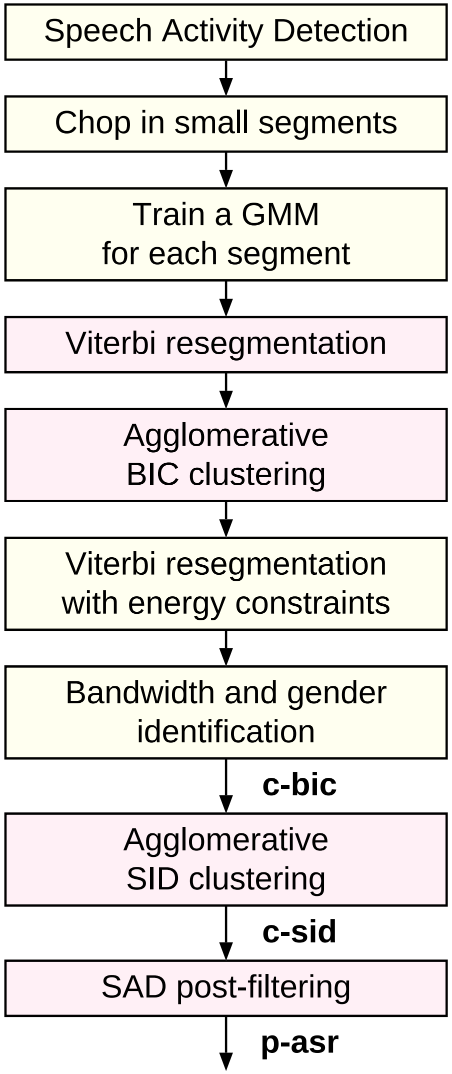
## Data type

- Broadcast News (BN)
- Lecture recordings: seminars and conferences

## Challenges of lecture / seminar data

- spontaneous speech with overlaps
- variations of microphone characteristics and positions for MDM condition
- crosstalk in IHM condition

# SPEAKER DIARIZATION SYSTEM FOR BN (1)




## Front-end

- 38 features: 12 MFCC + 12 $\Delta$ + 12 $\Delta\Delta$ + $\Delta$ E + $\Delta\Delta$ E

## Speech activity detection

- Viterbi decoding with 5 models: speech, music, speech over music, noise, silence GMMs (64 Gaussians)

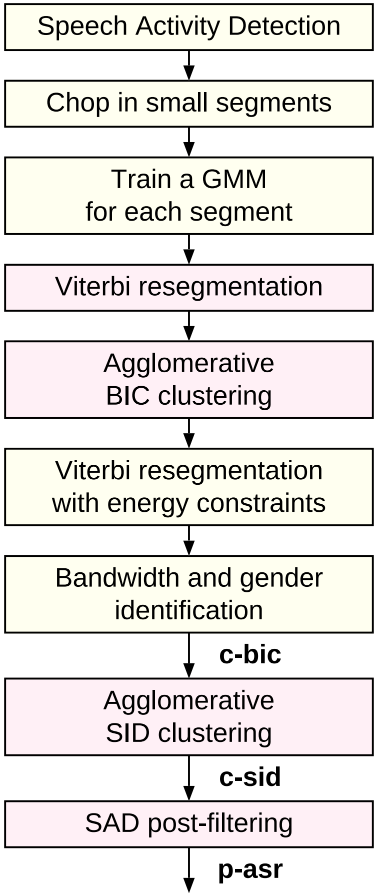
## Chop into small segments

- 2 sliding windows of 5 sec, local divergence measure

## GMM estimation for each segment

- 8-component GMM with diagonal covariance matrix per segment

# SPEAKER DIARIZATION SYSTEM FOR BN (2)




## BIC Agglomerative clustering

- Gaussian with full covariance matrix
- merge criterion

$$\Delta BIC = (n_i + n_j) log|\Sigma| - n_i log|\Sigma_i| - n_j log|\Sigma_j| - \lambda$$

with penalty

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N$$

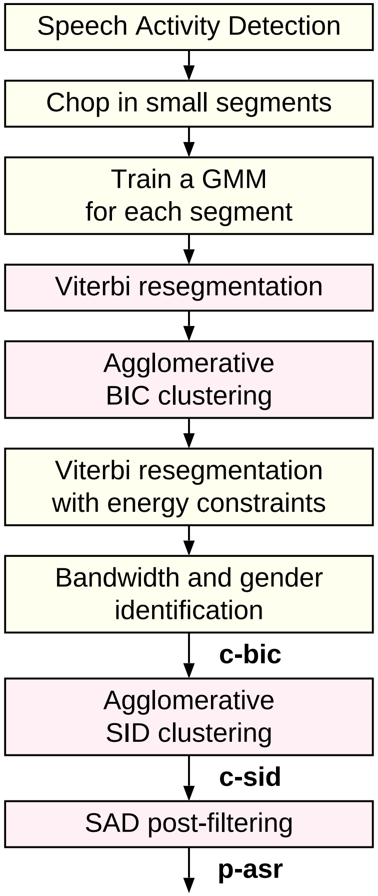
- stop criterion

$$\Delta BIC >= 0$$

## BIC penalty

- local: $N = n_i + n_j$
- global: $N = \Sigma_k n_k$

# SPEAKER DIARIZATION SYSTEM FOR BN (3)




## SID clustering

- 15 MFCC + $\Delta$ + $\Delta$ E, feature warping (Gaussian normalization)
- Universal Background Models (UBM) with 128 Gaussians (male/female, studio/telephone)
- MAP adaptation of matching UBM
- cross log-likelihood ratio between clusters $c_i$ and $c_j$

$$clr(c_i, c_j) = \frac{1}{n_i} log \frac{f(x_i|M_j)}{f(x_i|UBM)} + \frac{1}{n_j} log \frac{f(x_j|M_i)}{f(x_j|UBM)}$$

with $x_i$ the data from cluster $c_i$, $M_i$ the model for cluster $c_i$, $n_i$ the size of segment $x_i$

- threshold $\delta$

# SAD ADAPTED ON LECTURE DATA

## SAD on BN

- Viterbi decoding with 5 models trained on BN data: speech, noise, speech over music, pure music and silence GMMs

## Log-likelihood based SAD for lecture data

- GMMs for speech and non-speech (silence or noise)
- trained on far-field data (7 ISL seminars recorded in 2003)
- log-likelihood (llh) ratio between 2 models computed for each frame
- different prior probabilities for speech and non-speech
- transition points detected by the maxima of the mean of llh over smoothing window of 100 frames

# AUDIO INPUT FOR SPKR SYSTEM

| *dataset* | *condition* | *AIT* | *IBM* | *ITC* | *UKA* | *UPC* |
|---|---|---|---|---|---|---|
| dev | MDM | mic05 | Audio_17 | Table-1 | TableTop-1 | channel15 |
| eval | MDM | mic06 | Audio_17 | Table-2 | TableTop-1 | channel16 |
| eval | SDM | mic05 | Audio_19 | Table-1 | Table-2 | channel15 |

- MDM: single microphone signal randomly selected from available MDM channels and different from the channel of SDM
- SDM: single distant microphone signal defined by NIST
- MM3A: beam-formed multiple mark III microphone array data provided by Karlsruhe lab

# SAD EXPERIMENTS ON LECTURE DATA

- varied the acoustic data use to train the GMMs
- varied the number of GMMs
- viterbi decoding vs smoothed llh based SAD
- varied the number of Gaussians per mixture
- varied prior probabilities for speech/non-speech

# RESULTS ON DEV FOR MDM (1)

| *system* | *Missed speech(%)* | *False alarm speech(%)* | *Speaker error(%)* | *overlap SPKR Err.(%)* |
|---|---|---|---|---|
| vitbn | 18.2 | 3.0 | 9.0 | 30.19 |
| vitbn+mt | 19.3 | 2.9 | 8.7 | 30.96 |
| vitmt | 14.2 | 3.7 | 12.4 | 30.23 |
| gmtmt | 2.7 | 6.1 | 11.7 | 20.53 |

## Different SAD used in the speaker diarization system

- vitbn: Viterbi decoding using 5 GMMs(64 Gaussians) trained on BN data
- vitbn+mt: Viterbi decoding using GMMs trained on BN data plus GMMs (256 Gaussians) for speech and non-speech trained on lecture data
- vitmt: Viterbi decoding only using 2 GMMS trained on lecture data
- gmtmt: log-likelihood ratio based SAD with a prior probability of 0.2 for non-speech and 0.8 for speech
- BIC penalty weight $\boldsymbol{\lambda = 3.5}$, SID threshold $\boldsymbol{\delta = 0.5}$

# RESULTS ON DEV FOR MDM (2)

| *nb. Gaussians* | *Missed speech(%)* | *False alarm speech(%)* | *Speaker error(%)* | *overlap SPKR Err.(%)* |
|---|---|---|---|---|
| 64 | 9.5 | 4.0 | 11.0 | 24 |
| 128 | 9.5 | 3.7 | 11.0 | 24 |
| 256 | 7.8 | 4.2 | 11.0 | 23 |
| 512 | 7.7 | 4.2 | 11.1 | 23 |

## Models with varied number of Gaussians used in llh based SAD

- with a prior probability for non-speech and speech being $\mathbf{0.4 : 0.6}$
- BIC penalty weight $\boldsymbol{\lambda} = \mathbf{3.5}$
- SID threshold $\boldsymbol{\delta} = \mathbf{0.5}$
- no improvements above 256 Gaussians

# RESULTS ON DEV FOR MDM (3)

| *P(NS):P(S)* | *Missed speech(%)* | *False alarm speech(%)* | *Speaker error(%)* | *overlap SPKR Err.(%)* |
|---|---|---|---|---|
| 0.1:0.9 | 1.0 | 9.5 | 12.0 | 22.43 |
| 0.2:0.8 | 2.7 | 6.1 | 11.7 | 20.53 |
| 0.3:0.7 | 5.2 | 5.0 | 11.3 | 21.51 |
| 0.4:0.6 | 7.8 | 4.2 | 11.0 | 22.99 |

## Different prior probabilities for non-speech and speech

- using 256-component GMMs for speech and non-speech
- BIC penalty weight $\lambda = 3.5$
- SID threshold $\delta = 0.5$

# EVALUATION RESULTS

| *system* | *overlap SAD Err.(%)* | *overlap SPKR Err.(%)* | *non-overlap SPKR Err.(%)* |
|---|---|---|---|
| MDM_g256_p2-8 | 9.53 | 24.58 | 24.09 |
| SDM_g256_p2-8 | 12.09 | 26.21 | 25.79 |
| MM3A_g128_p4-6 | 12.85 | 27.11 | 26.57 |

## Configuration of SAD

- "g": number of Gaussians used in GMMs for speech and non-speech
- "px-y": prior probabilities for non-speech and speech being 0.x:0.y

## Configuration of SPKR

- BIC penalty weight $\boldsymbol{\lambda = 3.5}$
- SID threshold $\boldsymbol{\delta = 0.5}$

# CONCLUSIONS

## Speaker diarization system for lecture data

- log-likelihood based SAD reduced SAD error by 60% on dev for MDM
- overall diarization error on test data:
  24.6% for MDM and 27.1% for MM3A
- combination of BIC and SID clustering effective on both BN and lecture data
- blind speaker diarization: we hypothesize that the entire signal is speech from one speaker, (i.e. no non-speech)
  SAD error of 13% and SPKR error of 27% for MDM

## Future directions

- preprocessing step to combine all MDM channels to one single channel
- combination of SAD results on each MDM channels